RESEARCH PROPOSAL

"A Statistical Study on some Machine Learning Optimization Algorithms"

A synopsis submitted to

Swami Ramanand Teerth Marathwada University,Nanded (M.S.)

For the degree of

Doctor of Philosophy

In

Statistics

Under the faculty of

Science and Technology

By

Ms. Pooja Sarjerao Shitole
M.Sc. (Statistics),
Phone No: 9158384381
Email ID: shitolepooja16@gmail.com

Under the guidance of

Dr. Aniket Avinash Muley
Associate Professor,
Department of Statistics,
School of Mathematical Sciences,
Swami Ramanand Teerth Marathwada University,
Nanded-431606 (M.S.) INDIA
Eamil ID:aniket.muley@gmail.com
Phone No:+917276114558

August 2023

Table of Contents

Sr. No.	Contents	Page No.
1	Introduction	1-2
2	Review of Literature and Development in the subject	2-3
3	Objectives of the Research	3
4	Methodology	3-4
5	Importance of Study	4-5
6	Significance of Study	6
6	Proposed Work Plan	6
7	References	7-8

1. Introduction

Machine Learning (ML) is one of the most popular sub-fields of Artificial Intelligence (AI). Nowadays, ML is used almost everywhere viz., Healthcare, Finance, Infrastructure, Marketing, Self-driving cars, recommendation systems, chat-bots, social sites, gaming, cyber security, and many more. It has several practical applications that drive the kind of real business results as time and money savings and that have the potential to dramatically impact the future. In particular, tremendous impact occurring within the customer care industry, whereby ML is allowing people to get things done more quickly and efficiently. Through Virtual Assistant solutions, ML automates tasks that would otherwise need to be performed by a live agent viz., changing a password or checking an account balance. This frees up valuable agent time that can be used to focus on the kind of customer care that humans perform best: high touch, complicated decision-making that is not as easily handled by a machine. Further, it improves the process by eliminating the decision of whether a request should be sent to a human or a machine: unique Adaptive Understanding technology, the machine learns to be aware of its limitations and bailout to humans when it has low confidence in providing the correct solution. Currently, ML is under the development phase, and many new technologies are continuously being added into it and are helpful in many aspects such as analyzing large chunks of data, data extractions, interpretations, etc. [1, 4-6, 12, 14, 18, 20, 23]

Some of the major areas in which ML is used as: [1, 7, 12, 14, 17, 22]

- **Healthcare:** ML is widely used in the healthcare industry. It helps healthcare researchers to analyze data points and suggest outcomes. Natural language processing helped to give accurate insights for better results of patients. Further, it has improved the treatment methods by analyzing external data on patients' conditions in terms of X-ray, Ultrasound, CT-scan, etc. NLP, medical imaging, and genetic information are key areas of machine learning that improve the diagnosis, detection, and prediction system in the healthcare sector.
- **Automation:** It is one of the significant applications of ML helps to make the system automated and that helps machines to perform repetitive tasks without human intervention. As a ML engineer and data scientist, you have the responsibilities to solve any given task multiple times with no errors and having the capability of performing iterative tasks in lesser time.
- **Banking and Finance:** ML is a subset of AI that uses statistical models to make accurate predictions. In the banking and finance sector, and it helped in fraud detection, portfolio management, risk management, chatbots, document analysis, high-frequency trading, mortgage underwriting, AML detection users can easily identify whether the email is useful or spam. This is also possible by machine learning algorithms such as Multi-Layer Perceptron, Decision tree**,** and Naïve Bayes classifier. Content filter, header filter, rules-based filter, permission filter, general blacklist filter, etc., are some important spam filters used by Google.
- **Self-driving cars:** ML plays a vital role in the manufacturing of self-driving cars. It uses an unsupervised learning method to train car models to detect people and objects while driving.

Tata and Tesla are the most popular car manufacturing companies working on self-driving cars. Hence, it is a big revolution in a technological era which is also done with the help of machine learning.

- **Credit card fraud detection:** Credit card frauds have become very easy targets for online hackers. As the culture of online/digital payments is increasing, the risk of credit/debit cards is parallel increasing. ML helps developers to detect and analyze frauds in online transactions. It develops a novel fraud detection method for Streaming Transaction Data, with an objective to analyze the past transaction details of the customers and extract the behavioral patterns. Further, cardholders are clustered into various categories with their transaction amount so that the behavioral pattern of the groups can be extracted respectively. Hence, credit card fraud detection is a novel approach using Aggregation Strategy and Feedback Mechanism of machine learning.
- **Stock Marketing and Trading:** ML also helps in the stock marketing and trading sector, where it uses historical trends or past experience for predicting the market risk. As share marketing is another name of marketing risk, machine learning reduces it to some extent and predicts data against marketing risk. Machine learning's **long short-term neural memory network** is used for the prediction of stock market trends.
- **Language Translation:** The use of ML can be seen in language translation. It uses the sequence-to-sequence learning algorithms for translating one language into other. Further, it also uses images recognition techniques to identify the text from one language to other. Similarly, Google's GNMT (Google Neural Machine Translation) provides this feature, which is a Neural Machine Learning that translates the text into our familiar language, and it is called automatic translation.

2. Review of Literature and Development in the subject.

Huang et al. [11]performed comparative study of the effects of the single model and integrated model of these machine learning algorithms in the application of regression problems, the advantages and disadvantages of each machine learning algorithm are studied. The performance of each ML algorithm in regression prediction is verified by simulation experiments on four different data sets. They showing the result research on several ML algorithms and their integration models have certain feasibility and rationality. Rajita et al. [21] investigates the effectiveness of Cuckoo search optimization (CSO), particle swarm optimization (PSO), ant colony optimization (ACO), jellyfish search optimization (JFO), and mayfly optimization (MFO) evolutionary algorithms in tuning the hyper-parameters of four ML models to achieve higher accuracy in the results. The comparative analysis of 20 combinations (five evolutionary algorithms and four ML models) showed that CSO improves average accuracy by 4.12% in all the ML models compared to PSO. Kai Guo et al. [9-10] investigated the effectiveness of Cuckoo search optimization (CSO), particle swarm optimization (PSO), ant colony optimization (ACO), jellyfish search optimization (JFO), and mayfly optimization (MFO) evolutionary algorithms in tuning the hyper-parameters of four ML

models to achieve higher accuracy in the results. The comparative analysis of these 20 combinations (five evolutionary algorithms and four ML models) showed that CSO improves average accuracy by 4.12% in all the machine learning models compared to PSO.

3. Objectives of the Research

Some of the proposed objectives of the study are:

1. To understand the various algorithms in the ML.
2. Try to explore the open source data with ML techniques.
3. Try to develop new ML optimization algorithm for different datasets.
4. Try to identify the new applications of the existing ML optimization algorithms.
5. Try to compare and optimize existing real life data with proposed model.

During the period of research many new objectives may be developed which willbe beneficial for the society.

4. Methodology:

In this study, the research can be carried out by analysis of existing machine learning predictive techniques on different datasets. The research requires various different kinds of databases, tools and techniques, machine learning algorithms, scientific languages, different software and packages.

Mainly Machine Learning techniques are classified into following three types: [1, 3, 9, 13, 22]

- **Supervised Machine Learning**

  In supervised learning the model is trained with labeled data then the algorithms predict the output from input data. The model becomes in-charge of data and tells the model what to predict. Classification, regression, k-nearest neighbors (KNN), Naive Bayes (NB), Support Vector machines (SVM) and Decision Trees (DT) algorithms are come under Supervised Machine Learning.

- **Unsupervised Machine Learning**

  In unsupervised learning the model is trained with unlabelled data then the algorithms learn to inherent patterns from the input data. In this the model has no idea about the data but tries to find grouping in data. Clustering and Association algorithms come under this type of Machine Learning.

- **Reinforcement Machine Learning**

  It involves the model to learn from its past errors and correct it in the next iteration. Inthis way model improves its results and increase efficiency.

➢ **Software and Packages**

Some of the most used packages for ML

- Python
- R
- WEKA
- MATLAB
- MOJO

The research methodology uses above mentioned different software packages to analyze and compare the predictive statistical techniques on different datasets.

5. Importance of Study/Society application:

- **National and International Status and some diversified applications of Machine learning**:

Patro et al. [19] studied large data sets of fiducially-based features from ECG are extracted in an effective way. Moreover, the statistical learning method (LASSO) has been proposed for selecting appropriate features for classification. Furthermore, the refined features are subjected to popular machine learning classification algorithms for the evaluation of recognition system performance. Rajita et al. [21] investigated the effectiveness of Cuckoo search optimization (CSO), particle swarm optimization (PSO), ant colony optimization (ACO), jellyfish search optimization (JFO), and mayfly optimization (MFO) evolutionary algorithms in tuning the hyper-parameters of four ML models to achieve higher accuracy in the results. The accuracy of the neural network model improved after tuning its hyper-parameters using CSO. Huang et al. [11] applied several common ML algorithms: BP neural network, support vector machine and extreme learning machine to the regression problem. Further, 2 individual synthetic data sets and 2 real data sets were used from different application fields. The model uses three performance evaluation indicators: mean square error (MSE), mean absolute percentage error (MAPE) and square correlation coefficient (R^2). The experimental results of the integrated model show that the integrated model was found to superior to the single model using ML algorithm. Through the performance comparison experiments, several ML algorithms and their integration models studied. Guo et al. [9-10] discarded the traditional two-stage modeling approach and proposed an integrated chaotic time series prediction model based on EELM and differential evolution (DE). They have used EELM as an efficient and robust learning algorithm to model chaotic time series. Meanwhile, a modified DE algorithm is employed to search for the optimal parameter combination of embedding dimension, time delay of chaotic time series and the hidden nodes number of EELM simultaneously based on a hybrid validation criterion gives global optimum or approximate optimum which enhances the prediction performances of the model, but also tactfully bypasses the problem that traditional prediction models depend too much on the phase space reconstruction. Mudassir et al. [17] demonstrated high-performance machine learning-based classification and regression models for predicting Bitcoin price movements and prices in short and medium terms. They found that, their proposed models are feasible and gave high

performance forecast. Elen and Avuclu [8] proposed a new hybrid model has been developed to classify EEG and ECG signals. Their proposed hybrid machine learning model is consisting of four stages. In the experimental studies, they observed that the proposed hybrid method gives better results than the normal classification process. Makkhan et al. [16] developed a wavelet-ANFIS conjugation model and a wavelet-ARIMA coupled model along with the timeseries ARIMA model for the prediction of black carbon concentration over the Raniganj, Jharia, and Bokaro coal mines of India, by considering a long term data obtained by NASA and compared the results for determining the best prediction model. Further, validity of the results is tested with the help of error measures viz., RMSE, MSE, MAPE, MAE, and relative error. They observed that the Wavelet-ANFIS conjugation approach outperforms the wavelet-ARIMA coupled approach and the simple time-series ARIMA model. Li and Wu [15] dealt with three different datasets: 1) the heat load prediction data of building energy, mainly predict the heat load of HVA system according to the compactness, surface area, height and direction of buildings; 2) The cold load prediction data of building energy, mainly predict the cold load of HVA system according to the compaction rate, surface area, height and direction of buildings; Turbine attenuation state prediction data, mainly predict the decay state coefficient of ship turbine according to the lever position and ship speed; 3) Concrete compressive strength data; mainly predict the compressive strength of concrete according to cement component, high slag, fly ash, water and other components.

Chang et al.[3] used the machine learning algorithm for the planning and prediction of product R&D. Firstly, the prediction model of a product's success rate is developed by using the random tree algorithm, the bagging algorithm and the SMO algorithm. The prediction results of the three algorithms are integrated to judge the market's acceptance of the planned new products. Finally, the new products are improved, and after the improvements have been made, the three algorithms are used to make predictions. They observed that, the acceptance of new products by customers has increased significantly. Boutaba et al. [2] discussed the advances made in the application of ML in networking. They have focused on traffic engineering, performance optimization and network security. Sarker [22] dealt with different machine learning techniques and their applicability in various real-world application domains viz., cyber-security systems, smart cities, healthcare, e-commerce, agriculture, and many more. Also, emphasize on the challenges and potential research directions of their study. Kumar et al.[13] investigated and predicted the future environmental circumstances of a COVID-19 to minimize its effects using artificial intelligence techniques. The experimental investigation of COVID-19 instances has been performed in ten countries. The country dataset includes: India, the United States, Russia, Argentina, Brazil, Colombia, Italy, Turkey, Germany, and France using machine learning, deep learning, and time series models. The confirmed, deceased, and recovered datasets of Novel COVID-19 cases were considered from the Kaggle COVID dataset repository. Out of different machine learning, deep learning, and time series models, Random Forest Regressor, Facebook Prophet, and Stacked LSTM outperformed to predict the best results for COVID-19 instances with the lowest root mean square and highest R^2 score values.

6. Significance of the study

Data is the key term in all the businesses. It is important to have data for better decisions to keeping up with competition and falling further behind with other competitor industries. ML can be helpful to indentify the success key of corporate and customer data that keep a company ahead. In today's world, ML is important to identify the trends in customer behavior and operational business patterns, as well as supports the development of new products. Many of todays leading companies' viz., Facebook, Google, and Uber etc., make ML a key part of their operations. Also, it is significantly used in everyday scenarios viz., financial services fraud detection, retail purchase predictions, and online customer support interactions. In AI for applications viz., natural language processing (NLP) and computer vision (CV) are helpful in industries like financial services, healthcare, and automotive accelerate innovation, improve customer experience, and reduce costs. NLP and CV provide a valuable link between humans and robots: NLP helps computer programs understand human speech, and CV applies machine learning models to images, and is perfectly useful for the whole things from selfie filters to medical imaging data set.[1, 3, 6, 13, 14, 20, 22]

7. Proposed work Plan/ Formulation and Structure of Study:

- **Year wise Plan of work and targets to be achieved**
- **First Year**

1. To collect and study review of literature of ML techniques.
2. Collection and study of different ML data sets.
3. Study of different ML Software.

- **Second Year**

1. Try to develop statistical model for ML.
2. Selection of the standard datasets to study statistical machine learning technique and further it will be compare with the proposed algorithms
3. Detailed study and collaborative analysis of the results obtained.
4. The comparative study of the use of different ML software for the same data sets.

- **Third Year**

1. The publication of the findings of the research.
2. Publication of papers in various national and international journals and conferences.
3. Thesis writing.

8. References:

1. Abadeh, M. N., & Mirzaie, M. (2023). A differential machine learning approach for trust prediction in signed social networks. The Journal of Supercomputing, 1-24.
2. Boutaba, R., Salahuddin, M. A., Limam, N., Ayoubi, S., Shahriar, N., Estrada-Solano, F., & Caicedo, O. M. (2018). A comprehensive survey on machine learning for networking: evolution, applications and research

opportunities. Journal of Internet Services and Applications, 9(1), 1-99.

3. Chang, Y. T., Yang, H. R., & Chen, C. M. (2022). Analysis on improving the application of machine learning in product development. The Journal of Supercomputing, 78(10), 12435- 12460.
4. Cisneros, D., Gong, Y., Yadav, R., Hazra, A., & Huser, R. (2023). A combined statistical and machine learning approach for spatial prediction of extreme wildfire frequencies and sizes. Extremes, 1-30.
5. Cooper, G. (2023). Examining science education in chatgpt: An exploratory study of generative artificial intelligence. Journal of Science Education and Technology, 1-9.
6. Dhelim, S., Chen, L., Ning, H., & Nugent, C. (2023). Artificial intelligence for suicide assessment using Audiovisual Cues: a review. Artificial Intelligence Review, 56(6), 5591- 5618.
7. Doppala, B. P., Bhattacharyya, D., Chakkravarthy, M., & Kim, T. H. (2021). A hybrid machine learning approach to identify coronary diseases using feature selection mechanism on heart disease dataset. Distributed and Parallel Databases, 1-20.
8. Elen, A., & Avuçlu, E. (2022). A hybrid machine learning model for classifying time series. Neural Computing and Applications, 1-19.
9. Guo, K., Yang, M., & Zhu, H. (2020). Application research of improved genetic algorithm based on machine learning in production scheduling. Neural Computing and Applications, 32,1857-1868.
10. Guo, W., Xu, T., & Lu, Z. (2016). An integrated chaotic time series prediction model basedon efficient extreme learning machine and differential evolution. Neural Computing andApplications, 27, 883-898.
11. Huang, J. C., Ko, K. M., Shu, M. H., & Hsu, B. M. (2020). Application and comparison of several machine learning algorithms and their integration models in regression problems. Neural Computing and Applications, 32, 5461-5469.
12. Konar, Sushant, Nitin Auluck, Rajarajan Ganesan, Atul Kumar Goyal, Tarunpreet Kaur, Mansi Sahi, Tanvir Samra, Shyam Kumar Singh Thingnam, and Goverdhan Dutt Puri. (2022).A non-linear time series based artificial intelligence model to predict outcome in cardiac surgery. Health and Technology, 1-13.
13. Kumar, Y., Koul, A., Kaur, S., & Hu, Y. C. (2022). Machine Learning and Deep Learning Based Time Series Prediction and Forccasting of Ten Nations' COVID-19 Pandemic. SN Computer Science, 4(1), 91.
14. Kwaghtyo, D. K., & Eke, C. I. (2022). Smart farning prediction models for precision agriculture: a comprehensive survey. Artificial Intelligence Review, 1-44.
15. Li, J., & Wu, Y. (2022). Improved Sparrow Search Algorithm with the Extreme Learning Machine and Its Application for Prediction. Neural Processing Letters, 54(5), 4189.4209.
16. Makkhan, S. J. S., Singh, S., Parmar, K. S., Kaushal, S., & Soni, K. (2023). Comparison of hybrid machine learning model for the analysis of black carbon in air around the major coal mines of India. Neural Computing and Applications, 35(4), 3449-346
17. Mudassir, M., Bennbaia, S., Unal, D., & Hammoudeh, M. (2020). Time-series forecasting of Bitcoin prices using high-dimensional features: a machine learning approach. Neural computing and applications, 1-15.
18. Pan, Guangfei, Feiyang Wang, Chunlei Shang, Honghui Wu. Guilin Wu, Junheng Gao, Shuize Wang, Zhijun Gao, Xiaoye Zhou, and Xinping Mao (2023). Advances in machine learning-and artificial intelligence-assisted material design of steels. International Journal of Minerals, Metallurgy and Materials, 30(6), 1003-1024.
19. Patro, K. K., Reddi, S. P. R., Khalelulla, S. E., Rajesh Kumar, P., & Shankar, K. (2020). ECG data optimization for biometric human recognition using statistical distributed machine learning algorithm. The Journal of Supercomputing, 76, 858-875.
20. Qiu, J., Jammalamadaka, S. R., & Ning, N, (2020). Multivariate time series analysis from a Bayesian machine

learning perspective. Annals of Mathematics and Artificial Intelligence, 88(10), 1061-1082.

21. Rajita, B. S. A. S., Tarigopula, P., Ramineni, P., Sharma, A., & Panda, S. (2023). Application of Evolutionary Algorithms in Social Networks: A Comparative Machine Learning Perspective. New Generation Computing, 1-44.
22. Sarker, I. H. (2021). Machine learning: Algorithms, real-world applications and research directions. SN computer science, 2(3), 160.
23. Srinivasan, R., & Subalalitha, C. N. (2021). Sentimental analysis from imbalanced code mixed data using machine learning approaches. Distributed and Parallel Databases, 1-16.

Ms.Pooja Sarjerao Shitole
Research Scholar

Dr.Aniket A. Muley
Research Guide

Place: Nanded
Date: 25/08/2023